Applying the TRMS-PE measure of jerkiness to one particular codec has illustrated several important insights into the operation of VTC/VT. First, the TRMS-PE jerkiness measure is very stable; in fact, the small variation in speed between consecutive swings of the ball shows up as a small variation in TRMS-PE, as expected. Second, the jerkiness, or temporal update of the codec, may not vary with code rate (as in Figure 20). The particular codec tested here achieved bit reduction by degrading the spatial resolution of scenes and not the frequency of update. For the particular codec tested, the jerkiness was due to omission of every other frame, regardless of operating bit rate. This simple result would not necessarily have been obtained for codecs that use more sophisticated coding/decoding methods. Other codec algorithms for attaining less jerkiness might trade spatial resolution for more or less temporal positioning accuracy.

2.8 Jerkiness Feature Using Difference Image

The TRMS-PE measure of jerkiness cannot easily be applied to arbitrary video scenes. Section 2.8.1 proposes a measure of jerkiness that can be applied to any video scene. The genesis of this new measure of jerkiness occurred when observing codec input and output video that had been aligned using the single-frame temporal alignment method discussed earlier (as in Figure 14). If one were to compute the difference images of the input and the output video, image pairs that contained no positioning errors (second and third images of each row in Figure 14) would yield smaller difference errors than image pairs that contained positioning errors (first and fourth images of each row in Figure 14). As a function of time, the total composite difference error of a moving object would be composed of two components. One component represents errors due to blurring or distortion of the object. The other component represents errors due to incorrect positioning of the object.

Section 2.8.1 presents a method for extraction of three new features. One of the features will be shown to be intimately related to jerkiness. The other two features represent the average distortion of the output video due to jerkiness and spatial blurring. The exact feature extraction technique and sample VTC/VT results are discussed in detail next.

2.8.1 Feature Extraction Technique

The features are extracted from the undistorted input and distorted output sampled video. The feature extraction technique is computationally efficient and possesses many of the other desirable properties of features that were previously mentioned. The features are extracted from the standard deviation of the difference images (input image minus output image), where the input and output video has been time aligned using the single-frame temporal alignment method. The standard deviation of the difference image is used, instead of the mean or root mean square, because the standard deviation is insensitive to gray level shifts in the sampled video. The exact feature extraction method follows:

1. Video alignment

   Single-frame temporal alignment of the input and output video is performed.

2. Difference image

   For each aligned video image pair, a difference image is formed by subtracting the output image from the input image.

3. Standard deviation of the difference image (SD-DI)

   The standard deviation of each difference image (SD-DI), from step 2 above, is computed as the square root of (the summation of the squares of the image pixel values divided by the total number of pixels, minus the square of the mean of the difference image). Here, the mean of the difference image is computed as the summation of the image pixel values divided by the total number of pixels. See equation 16 in Appendix A for a mathematical definition of SD-DI.

4. Feature computation

   From the time history of SD-DI, from step 3 above, the following three features are computed:

   a. The temporal mean of SD-DI (TM-SD-DI)

      TM-SD-DI is computed as the summation of the SD-DI values divided by the total number of SD-DI values. TM-

SD-DI is primarily related to the average distortion caused by spatial blurring and jerkiness. See equation 17 in Appendix A for a mathematical definition of TM-SD-DI.

b. The temporal standard deviation of SD-DI (TSD-SD-DI)

TSD-SD-DI is computed as the square root of (the summation of the squares of the SD-DI values divided by the total number of SD-DI values, minus the square of TM-SD-DI). This estimate of the standard deviation of the population of SD-DI time samples is asymptotically unbiased for a large number of SD-DI values. An alternate method of computing TSD-SD-DI that is unbiased for a small number of SD-DI values may be used instead (see, for example, Crow et al., 1960). TSD-SD-DI is primarily related to jerkiness. See equation 18 in Appendix A for a mathematical definition of TSD-SD-DI.

c. The temporal root mean square of SD-DI (TRMS-SD-DI)

TRMS-SD-DI is computed as the square root of (the summation of the squares of the SD-DI values divided by the total number of SD-DI values). TRMS-SD-DI is related to the total distortion caused by spatial blurring and jerkiness. See equation 19 in Appendix A for a mathematical definition of TRMS-SD-DI.

To compute the amount of distortion in the difference images with respect to the input images, the SD-DI values could be normalized by the standard deviation of the undistorted input video. Alternatively, normalized features could be obtained by dividing TM-SD-SD, TSD-SD-DI, and TRMS-SD-DI by the temporal mean of the standard deviation of the undistorted input video. Thus, normalized features closer to zero will represent smaller distortions while normalized features closer to one will represent larger distortions.

2.8.2 Sample VTC/VT Results

The VTC/VT imagery of Figure 7 was processed to extract the TM-SD-DI, TSD-SD-DI, and TRMS-SD-DI features as described above. The difference images were obtained by subtracting the output images (rows two, three, and four of Figure 7) from the single-frame temporally aligned input images (row one of Figure 7). Figure 21 shows the resulting difference images, where rows one, two, and three of Figure 7 correspond to codec bit rates of DS1, 1/2 DS1, and 1/4 DS1, respectively. For display purposes only, the difference images of Figure 21 have been scaled such that gray (intensity of 128) represents no error, black (intensities from 0 to 127) represents negative error, and white (intensities from 129 to 255) represents positive error. Note that images 1 and 2 (from left to right) for codec bit rates DS1 and 1/4 DS1 contain small errors, while images 3 and 4 for a codec bit rate of 1/2 DS1 contain small errors. The particular codec under test achieved some of its data compression by discarding every other input frame and repeating every other decoded output frame (one frame represents two images in Figure 7 since the images were grabbed for each field increment of the video recorder, and there are two fields for each NTSC frame). Since the input video was injected asynchronously for each of the three codec bit rates, there was no guarantee that the same frames would be discarded for all bit rates. In Figure 14, the same video frames were discarded for all bit rates (by chance) while in Figure 7 the same frames were discarded for the DS1 and 1/4 DS1 bit rates but different frames were discarded for the 1/2 DS1 bit rate. Thus, care should be taken to process a sufficiently long time sequence of images when extracting the TM-SD-DI, TSD-SD-DI, and TRMS-SD-DI features. Otherwise, inaccurate results may be obtained, particularly for codecs that discard a large number of video frames.

The first two difference images for rate DS1 in Figure 21 contain errors due to blurring. The third and fourth difference images contain errors due to blurring and jerkiness. Examining Figure 7 closely, each image in the NTSC video scene (top row) is unique and shows steady motion of the report crossing the man's face. Meanwhile, the third and fourth codec output images for rate DS1 (second row) are the same as the first and second codec output images, respectively. One can see from Figure 7 that the codec is performing frame repetition. Thus, on the repeated

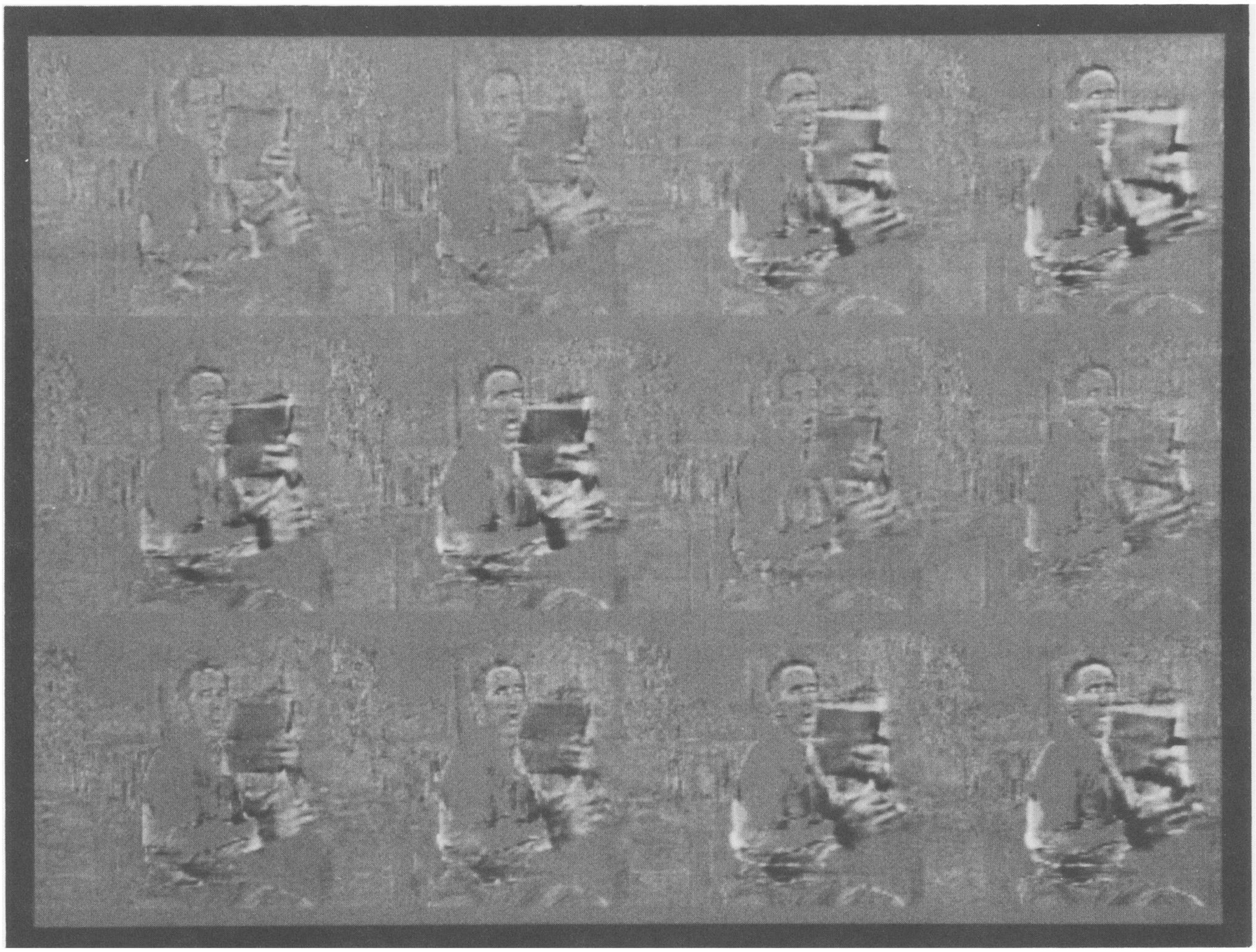

Figure 21. Difference images for VTC/VT imagery of Figure 7. Codec bit rates of DS1 (top row), 1/2 DS1 (second row), and 1/4 DS1 (bottom row).

frame (consisting of images 3 and 4 in Figure 7), large difference errors are obtained and these errors are due to jerkiness in the codec output.

Figure 22 is a graph of the time history of the SD-DI values for the images in Figure 21. The SD-DI values for the first four fields in Figure 22 were calculated from the first four difference images in Figure 21. The ball test scenes (Figures 14 and 15) and the man test scene (Figure 7) were obtained from the same codec. The time history of SD-DI very much resembles the codec output ball position errors (compare with the output ball position error with respect to the true input ball position in Figures 18).

Table 9 presents the computation of the unnormalized TM-SD-DI, TSD-SD-DI, and TRMS-SD-DI features for the eight fields of Figure 22 (for reference, the temporal mean of the standard deviation of the undistorted input video was 77.6). The temporal mean of SD-DI (TM-SD-DI) and the temporal root mean square of SD-DI (TRMS-SD-DI) represents the average and total distortion due to blurring and jerkiness. The temporal standard deviation of SD-DI (TSD-SD-DI) represents the extent of the variation of SD-DI about its mean. More jerky motion will result in larger values of TSD-SD-DI. Curiously, from Table 9, TSD-SD-DI increases slightly with increasing bit rate. This contradicts the earlier TRMS-PE measure of jerkiness which showed that jerkiness was the same for all bit rates (see Figure 20). An explanation for the phenomenon is as follows. The added spatial blurring for low bit rates versus higher bit rates tends to raise the SD-DI curve (increasing TM-SD-DI in Table 9). In the raised SD-DI curve, smaller increases in difference errors due to positioning are obtained, and this results in a flattening of the SD-DI curve (decreasing TSD-SD-DI in Table 9). Subjectively, the TSD-SD-DI measure of jerkiness may be more accurate than the TRMS-PE measure of jerkiness because added spatial blurring tends to reduce the effect of jerkiness. If the object is badly blurred, one cannot tell if the motion is jerky. If the object is focused, one readily notices jerky motion.

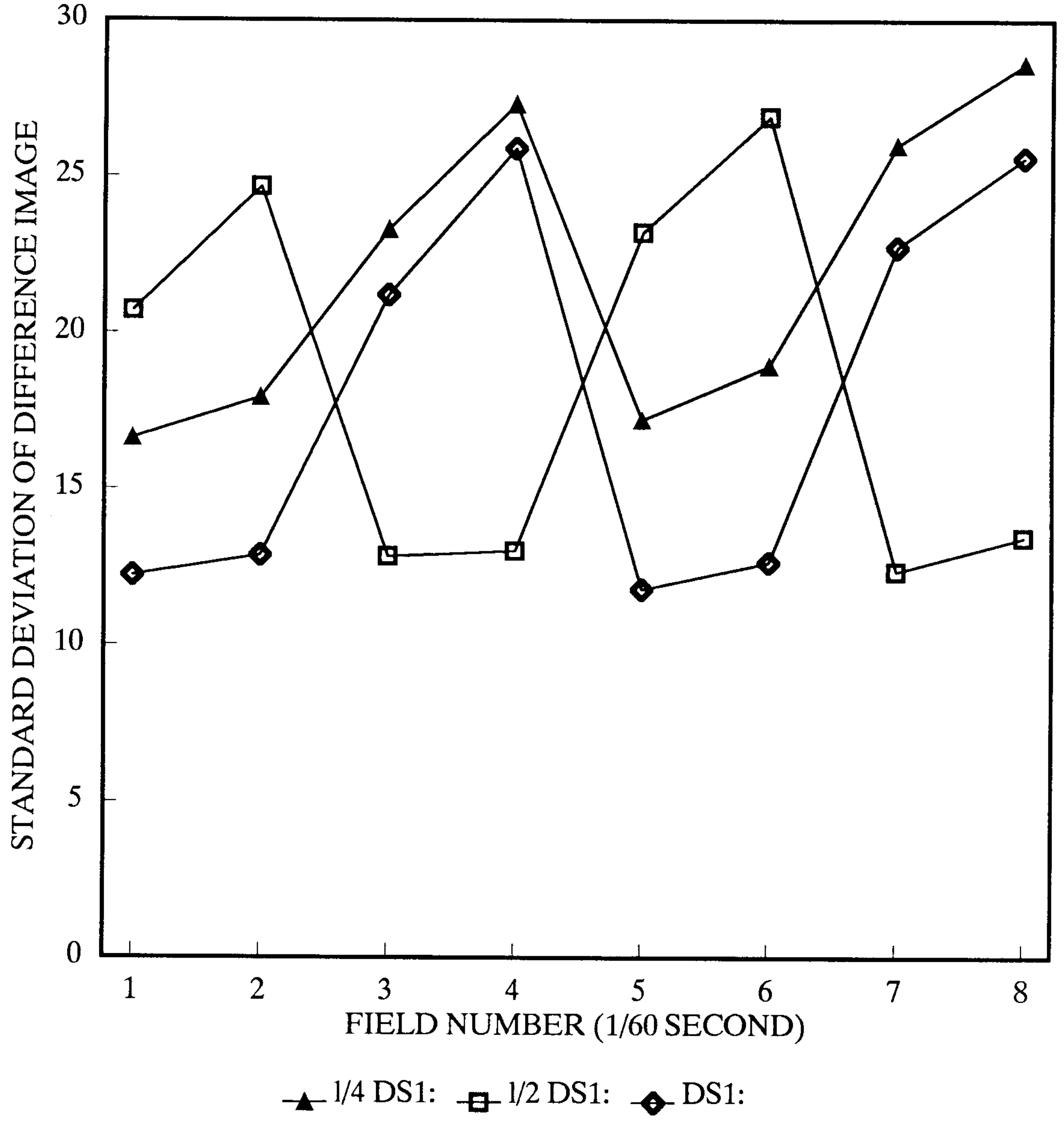

Figure 22. Time history of SD-DI for the difference images of Figure 21. The first four field numbers correspond to the four images in Figure 21.

Table 9. Summary Of SD-DI Features For Figure 22

Scene	TM-SD-DI	TSD-SD-DI	TRMS-SD-DI
DS1	18.1	5.9	19.1
1/2 DS1	18.4	5.7	19.3
1/4 DS1	22.0	4.6	22.5

3. CONCLUSIONS AND RECOMMENDATIONS

Objective feature extraction techniques have been presented that measure the predominant artifacts present in digitally transmitted video systems. Among these artifacts are blurring/smearing, blocking, edge busyness, image persistence, and jerkiness. Features are extracted from the digitized video imagery that reflect degradations perceived by the viewer. The features are sensitive to the type of video being transmitted which is important since the performance of digital codecs depend strongly on the type of video being transmitted. In addition, the features possess many of the desirable properties that humans also possess, including the potential adaptability to focus attention on local disturbances in the video. Thus, the features are expected to correlate strongly with subjective quality ratings.

Depending upon the feature one wishes to extract, the method for temporally aligning the input and distorted output video frames varies. Spatial blurring and jerkiness measures have been presented that do not require the input and output video scenes to be aligned. Other measures, such as edge busyness, blocking, image persistence, and jerkiness for natural motion scenes, require some form of temporal alignment. Two possible methods of temporally aligning the input and output video were presented. The computational requirements of the proposed features varied. However, these computational requirements appear reasonable for modern digital signal processing systems.

Spatial blurring features were presented that relate to the sharpness of the edges in the video imagery. These spatial blurring features appear to be applicable to many types of video imagery, including natural scenes. Blocking, edge busyness, and image persistence were shown to be forms of false edge energy appearing in the output